USING SINGLE-COUNTY VS. MULTI-COUNTY AS A PRIMARY SAMPLING UNIT

James Hartman, Randall Parmer, Hertz Huang, Dennis Schwanz, U.S. Census Bureau

I. INTRODUCTION

This paper presents the results of the primary sampling unit (PSU) definition research that was undertaken during the redesign of the nationally-based demographic surveys conducted by the Census Bureau. Specifically, these surveys were the American Housing Survey (AHS), which is conducted for the Department of Housing and Urban Development (HUD) to measure the level of and change in housing inventory characteristics; the National Crime Survey (NCS), which is conducted for the Department of Justice to measure the level of and change in victimization characteristics; the Health Interview Survey (HIS), which is conducted for the National Center for Health Statistics to measure the level of personal and household health characteristics; and the Survey of Income and Program Participation (SIPP), which is conducted and sponsored by the Census Bureau to measure the level of and change in income and income-transfer service program participation characteristics. Currently, all these surveys use multi-county PSUs; the research described herein was done to determine if the use of single-county PSUs would prove to be superior. Due to cost constraints, we restricted our analysis to AHS and HIS and used results from a previous study to make inferences for NCS and SIPP.

The method for evaluating the use of single-county vs. multi-county PSUs was based on a comparision of total variances and total costs that would result from each PSU definition. Since multi-county PSUs would tend to be more homogenous between PSUs than single-county PSUs, the between-PSU and total variances would be lower for the multi-county designs. Since the same number of cases would be interviewed under either PSU definition, the smaller land area associated with single-county PSUs would be expected to result in lower travel mileage and travel time during the interviewing. As a result, the travel costs and total costs would be lower for the single-county PSU design. Consequently the net variance-cost effect was the evaluation criterion for choosing between single and multi-county PSUs.

The rest of the paper consists of four sections with the first (Section II) describing the methodology utilized for determining the resultant variances for each type of PSU definition. The next section (Section III) describes the methodology utilized for determining resultant costs for each type of PSU definition. The next section (Section IV) presents results of the comparison of total and between variances for each PSU definition, the results of the comparison of the costs for each PSU definition, the results of the comparison of the deterioration of the two PSU designs, and the cost-variance comparison for each PSU definition. The deterioration aspect is described in more detail in the section describing the methodology used in determining the variances. The final section (Section V) presents the conclusions of the PSU definition research.

II. METHODOLOGY OF VARIANCE COMPARISON

To evaluate the effect of the two PSU definitions on the variances we first investigated its effects on variances of current data by stratifying and evaluating variances on the same year of an empirical data set. We then studied the increase in variances due to the deterioration of the stratification over time. It is possible that single-county PSUs may be less desirable than multi-county PSUs in 1980 based on a stratification of 1980 data. However, ten years from now single-county may prove superior based on the same stratification. This will occur if the multi-county variances increase significantly faster than single-county variances. We used a stratification based on 1980 data and variances based on 1970 and 1980 data to test this hypothesis.

We restricted our analysis to the case where both PSU definitions have the same areas defined as self-representing (SR) PSUs. These areas will be included in the sample with certainty in both PSU definitions. The main reason for this is because it is often desirable to keep all the counties in a large SMSA intact and the larger SMSAs are generally the SR PSUs. Downgrading small counties to nonself-representing (NSR) PSUs could yield different results.

Two files, both containing 1970 and 1980 census data from the South census region, were used. One file contained data for all counties in the South census region and the other contained data for all the multi-county PSUs in the South census region. The South was chosen because we believe that it is a fairly good region to represent the rest of the country. It has many rural PSUs which may be similar to rural PSUs in the West and Midwest regions; it has large urban areas, such as Atlanta, Miami, and Washington, D.C., which may be similar to PSUs in the Northeast region. By using only the South region, we were able to reduce the number of computer runs and thus reduce the cost.

Variables that were considered to be highly correlated with the objectives of these surveys were used as stratification variables. Other variables of importance were used as variance evaluation variables. In addition each stratification variable was given a weight based on its importance for a given survey. The evaluation and stratification variables for AHS and HIS are given in the table at the end of this paper.

For the stratification operation, we used a modified version of the Friedman-Rubin clustering algorithm (Reference 2). This stratification algorithm basically consists of three parts: the hill climbing pass, the exchange pass, and the size adjustment procedure. A modified Friedman-Rubin hill climbing procedure comprises a major portion of the stratification algorithm. In the modified procedure, PSUs are moved one at a time from one stratum to another in an attempt to reduce the between-PSU variance. The exchange pass also attempts to reduce the between-PSU variation by selecting pairs of PSUs from different strata and interchanging them. The size adjustment procedure creates strata which satisfy the strata

size constraints.

First, the Friedman-Rubin clustering algorithm was used to stratify 1980 census data. Once the optimum stratification was reached, the between-PSU within-stratum variances (to be referred to as between-PSU variances) were calculated for the items listed at the end of this paper. The items which have both 1970 and 1980 data were used as our evaluation variables. The 1980 data were used to evaluate the effects of the two PSU definitions on current data. The 1970 data, with the stratification based on 1980 data, were used along with the 1980 data to evaluate the time deteriorating effects for the two designs.

The ratio of single-county between-PSU variance to multi-county between-PSU variance was used in the evaluation of variances of the two types of PSU designs. A composite index, which is calculated by averaging these ratios of variances over all items, was used as a measure in the evaluation. The reason for using a ratio of the single-county variance relative to the multi-county variance is to eliminate the scale effect. This will permit small estimates to have the same importance as large estimates in the composite index.

As mentioned in Section I, we had intended to use total variances of the evaluation variables for comparison. However, at the time of the research, files that were needed for evaluating the effect on within-PSU variance were not available. We had to use results from another study (Reference 3) which showed that the within-county variances are about the same for counties of different sizes. Even though we used results from larger counties to represent multi-county PSUs, we were fully aware that larger counties could be different from multi-county PSUs with respect to within-PSU variance. Unfortunately, until we are able to examine the within-PSU variance for multi-county PSUs, we can only assume that the within-PSU variances are about the same for single-county and multi-county PSUs. Since systematic sampling is to be used for within-PSU sample selection, we believe that this assumption is reasonable. Therefore, the total variances were estimated from only the between-PSU variance component. Based on data from the current survey designs, the between-PSU variance for multi-county PSUs is believed to be between 10 percent and 25 percent of the total variance. Total multi-county variances were calculated for 4 different levels of between-PSU variance using the following formula:

$$\frac{\text{Multi-County Between-PSU Variance}}{\text{\% of Total Variance Attributable to Between-PSU Variance}}$$

The within-PSU variance for both PSU designs is equal to the total multi-county variance minus the multi-county between-PSU variance.

III. METHODOLOGY OF COST COMPARISON

The average savings from using a single-county design instead of a multi-county design were calculated using the following formula:

$$1 - \frac{\text{(Single-county design cost)}}{\text{(Multi-county design cost)}}$$

Total costs were estimated by adding total direct costs to overhead costs. Direct costs include costs due to mileage, travel time, interviewing, and questionnaire coding and editing. Overhead costs include recruiting, training, observation, and all office costs.

A. Direct Cost Model

The following cost model was used to estimate total direct costs:

(1) $C = hm(.20/mi) + h(T/60)(6.31/hr) + n(TI/60)(6.31/hr) + n(TO/60)(6.31/hr)$

where

- C = total direct costs per work assignment
- n^1 = number of cases per work assignment
- h^1 = average number of cases assigned for personal interview
- m = average distance traveled per household
- T = average travel time (minutes) per household
- TI = average interview length (minutes)
- TO = average time (minutes) spent on other direct cost activities like coding and editing

(.20/mi) represents the mileage reimbursement to interviewers, which is 20c per mile. (6.31/hr) represents the wage rate paid to interviewers, which is $6.31 per hour. (These represent 1983 figures).

m is calculated using the formula:

(2) $m = [(\lambda_1 s_1 - \lambda_2)\, d_1 + 2\lambda_2 d_2]/h$

T is calculated by:

(3) $T = [\lambda_1 s_1 - \lambda_2)\, d_1 r_1 + 2\lambda_2 d_2 r_2 + (\lambda_1 s_1) d_3 r_3]/h$

where:

- λ_1 = average number of visits per segment
- λ_2 = average number of trips from home
- s_1 = number of segments per assignment
- d_1 = average distance from segment to segment
- d_2 = average distance from home to segment
- r_1 = rate of travel between segments
- r_2 = rate of travel from home to segment
- $d_3 r_3$ = average time spent traveling within a segment

$(\lambda_1 s_1 - \lambda_2) d_1 r_1$ represents travel between segments, $2\lambda_2 d_2 r_2$ represents travel from home to segment, and $(\lambda_1 s_1) d_3 r_3$ represents travel within segments.

Adjustments were made for single and multi-county PSU designs to account for the fact that the distance between segments increases as the size of the PSU increases. These adjustments were achieved by applying the ratio of the distances as calculated by:

(4) $d^{2/} = \frac{1}{2}\sqrt{A/s_1}$

where A = average area per work assignment
s_1 = number of segments per assignment
d = estimated average distance between segments

B. Application of the Cost Model to Surveys

The cost calculations were made for AHS and HIS. For AHS, estimates of T (travel time per household interviewed) and m (mileage per household interviewed) were made directly from field data for a multi-county design. Values of T and m for the single-county design were estimated by applying the ratio of d_1's (distance between segments) for the two designs, as calculated by formula (4), and multiplying by the values of T and m for the multi-county design. An adjustment was made in T since within segment travel doesn't change. These estimates were applied to the direct cost formula (formula 1). To obtain average costs for the single and multi county PSU designs, a weighted mean between SR and NSR PSUs was used. The weights were proportional to the measure of size for SR and NSR areas in AHS which are .5515 and .4485, respectively.

For HIS, we had the national cost parameters for formulae (2) and (3). In order to estimate these cost parameters for SR and NSR PSUs for both single and multi-county designs, we modified the cost parameters of five population density categories that were obtained from the NCS interviewers' records. We then determined what proportion of PSUs fell in each population density category, separately for SR and NSR PSUs in the South region for the two designs. We used the proportions as weights to obtain a weighted average over the population density categories, for SR and NSR PSUs, for each cost equation parameter. The values of T and m were calculated using formulae (2) and (3) and were then used to calculate total direct cost (formula (1)).

Average costs for the single and multi-county PSU designs were determined by calculating a weighted mean between SR and NSR PSUs, as was done for AHS. The weights are .3977 and .6023, respectively, for SR and NSR PSUs.

C. Estimation of Total Costs and Cost Savings

An estimate of overhead costs were added to the direct costs for each survey to obtain the total cost. These total cost estimates were compared for the single and multi county PSU designs to obtain an estimate of the percent of total cost savings for a single-county PSU design over a multi-county PSU design.

IV. RESULTS

Results were obtained for two surveys, AHS and HIS. AHS consists of 100 NSR strata in the South. HIS stratified SMSA and non-SMSA PSUs separately. It consists of 32 NSR strata; 15 SMSA and 17 non-SMSA. This section presents the results of the single versus multi-county comparisons of between-PSU variance, deterioration, total variance, costs, and cost and variance combined for AHS and HIS.

A. Between PSU Variance

The results from our study show that single-county between-PSU variances are about 37 percent higher than multi-county variances for 1980 data for AHS. For 1970 data based on a stratification using 1980 data, single-county variances are about 36 percent higher.

For HIS, results were obtained separately for SMSA and non-SMSA as well as combined. The results show that single-county variances are about 47 percent higher than multi-county variances for SMSA PSUs and only about 26 percent higher for non-SMSA PSUs for 1980 data. This seems to indicate that non-SMSA multi-county PSUs are more similar to single-county PSUs than SMSA multi-county PSUs. For 1970 data based on the stratification of 1980 data, single-county variances are about 65 percent higher for SMSA PSUs and about 31 percent higher for non-SMSA PSUs. Overall, single-county variances are about 30 percent higher for 1980 data and about 41 percent higher for 1970 data.

B. Deterioration of Between-PSU Variance

To evaluate the deterioration of a single-county PSU design versus a multi-county PSU design we computed the coefficient of variation (cv) for 1970 and 1980 for single and multi-county PSU designs and compared the increase of the single-county cv to the increase of the multi-county cv from 1980 to 1970. The results are summarized in Table IV.B., below. For AHS, the results show that the increase in cv's are about the same for both single and multi-county PSUs.

For HIS, the results show that for SMSA PSUs there is a slight difference with multi-county cv's increasing about 28 percent and single-county cv's increasing about 35 percent. For non-SMSA PSUs both single and multi-county cv's increase about 25 percent. Overall, multi-county cv's increase about 22 percent while single-county cv's increase about 25 percent.

TABLE IV.B. CVs OF SINGLE VS. MULTI-COUNTY PSUs

	1980 SINGLE-COUNTY	1970 SINGLE-COUNTY	1980 MULTI-COUNTY	1970 MULTI-COUNTY
AHS	2.50%	3.54%	2.17%	3.05%
HIS				
SMSA	9.54%	12.92%	8.54%	10.97%
NONSMSA	10.20%	12.74%	9.61%	11.98%
COMBINED	7.02%	8.81%	6.58%	8.00%

C. Total Variance

As was explained previously, we assumed that the within-PSU variances for the two PSU designs were the same. Assuming that the between-PSU variance is between 10 and 25 percent of the total variance as is believed

under the current multi-county PSU design, the total variance for the two PSU designs is given in Table IV.C., below, for AHS and HIS. All variances are given relative to the multi-county between-PSU variance for the appropriate survey.

TABLE IV.C. SINGLE VS. MULTI-COUNTY TOTAL VARIANCE RATIOS

	SINGLE-COUNTY	MULTI-COUNTY	SINGLE / MULTI
AHS			
BETWEEN-PSU VARIANCE	1.367	1.000	1.367
TOTAL VARIANCE FOR :			
10 % BETWEEN-PSU VARIANCE	10.367	10.000	1.037
15 % BETWEEN-PSU VARIANCE	7.034	6.667	1.055
20 % BETWEEN-PSU VARIANCE	5.367	5.000	1.073
25 % BETWEEN-PSU VARIANCE	4.367	4.000	1.092
HIS : SMSA + NONSMSA			
BETWEEN-PSU VARIANCE	1.299	1.000	1.299
TOTAL VARIANCE FOR :			
10 % BETWEEN-PSU VARIANCE	10.299	10.000	1.030
15 % BETWEEN-PSU VARIANCE	6.966	6.667	1.045
20 % BETWEEN-PSU VARIANCE	5.299	5.000	1.060
25 % BETWEEN-PSU VARIANCE	4.299	4.000	1.075

As can be seen from this table, if AHS has a 20 percent between-PSU variance, then a 37 percent difference in between-PSU variance results in just a 7.3 percent difference in total variance.

D. Cost

The comparison of total costs for single and multi-county PSU designs are shown in Table IV.D. for 4 current surveys. The results show a cost savings of 0.8-2.0 percent for a single-county over a multi-county design. The potential savings varies from survey to survey due mainly to the proportion of total costs that is due to travel.

TABLE IV.D. SINGLE VS. MULTI-COUNTY COST PER WORK ASSIGNMENT COMPARISON

SURVEY / DESIGN	COSTS: DIRECT	COSTS: OVERHEAD	COSTS: TOTAL	PERCENT SAVINGS
AHS / SINGLE	987.09	2546.34	3533.43	2.0%
AHS / MULTI	1060.73	2546.34	3607.07	-
HIS / SINGLE	186.19	1036.18	1222.37	0.8%
HIS / MULTI	196.55	1036.18	1232.73	-
NCS / SINGLE	284.41	749.03	1033.44	2.0%
NCS / MULTI	305.06	749.03	1054.09	-
SIPP / SINGLE	631.25	3053.68	3684.93	1.4%
SIPP / MULTI	685.45	3053.68	3739.13	-

E. Cost and Variances

To evaluate the predicted cost savings for a single-county PSU design versus the expected higher variances, we evaluated the following ratio:

$$\frac{V_{single} \times C_{single}}{V_{multi} \times C_{multi}}$$

where V denotes variance and C denotes cost.

Obviously, a ratio greater than one means that the multi-county design is more desirable. The results for AHS and HIS are given in Table IV.E., below. The results show that a multi-county PSU design is more desirable from a cost-variance standpoint.

In fact, for AHS a design with less than 5.5 percent between-PSU variance would be necessary for single-county to be more desirable than multi-county. For HIS, a design with less than 2.8 percent between-PSU variance would be necessary.

TABLE IV.E. SINGLE VS. MULTI-COUNTY COST-VARIANCE RATIOS

	PERCENT BETWEEN-PSU VARIANCE: 10	15	20	25
AHS				
VARIANCE	1.031	1.047	1.062	1.078
COST	0.980	0.980	0.980	0.980
VARIANCE X COST	1.010	1.026	1.041	1.056
HIS				
VARIANCE	1.028	1.042	1.056	1.070
COST	0.992	0.992	0.992	0.992
VARIANCE X COST	1.020	1.034	1.048	1.061

V. CONCLUSIONS

Based on our results, we conclude that multi-county PSUs are superior to single-county PSUs. For the two surveys which we tested, AHS and HIS, the cost savings for a single-county PSU design were not great enough to offset the larger variances.

We believe this will be true for SIPP and NCS, also. A previous study, which used principal components (a linear combination of variables which explain most of the variances of these variables)[3] as statification variables, was conducted for the south region for two different numbers of NSR Strata (146 and 54). In that study, a modified version of the Friedman-Rubin clustering algorithm (Reference 2) was used to stratify 1970 census data. This previous study showed that for two different numbers of NSR Strata (146 and 54) the single-county to multi-county between-PSU variance ratio was very similar. In that study, the stratification containing 146 NSR strata produced single-county between-PSU variances which were about 40 percent higher than multi-county variances. For the stratification of 54 NSR strata, single-county variances were

about 34 percent higher. Since the number of NSR strata does not seem to affect the between-PSU variances, we strongly believe that single-county between variances will be at least 20 percent higher than multi-county variances for both SIPP and NCS. This is based on the fact that AHS and HIS showed that single-county variances were more than 30 percent higher than multi-county variances. If this is true, then for a design with 10 percent between-PSU variance, the variance x cost for SIPP and NCS will be the following :

	SIPP	NCS
Variance	1.020	1.020
Cost	0.986	0.980
Variance x Cost	1.006	0.9996

In addition, a stratification of multi-county-PSUs does not appear to deteriorate faster than a stratification of single-county PSUs. AHS showed that single and multi-county PSUs deteriorated at about the same rate. HIS showed that, overall, a stratification of single-county PSUs deteriorated slightly faster than a stratification of multi-county PSUs.

While we believe these conclusions to be true, there were some assumptions which were made due to the limited availability of certain data. Possibly the most important assumption was that the within-PSU variance of single and multi-county PSUs was the same. While this may not always be true, we don't believe the difference will be great enough to alter our conclusions. No data was available to verify this assumption.

REFERENCES

1. Friedman, H.P. and Rubin, J. (1967). "On Some Invariant Criteria for Grouping Data." Journal of the American Statistical Association 62, PP. 1159-1178

2. Kostanich, D.L., et al. (1981). "Modification of Friedman-Rubin's Clustering Algorithm for Use in Stratified PPS Sampling." Proceedings of the Section on Survey Research Methods, American Statistical Association, PP. 285-290.

3. Wolters, C.L. (1984). "HU Stratification in the American Housing Survey." Presented at the 1984 ASA Convention.

FOOTNOTES

1 The difference between n and h represents the number of cases assigned for telephone interview.

2 The correct formula is $d = 1/2 \sum_i \left(\pi_i \sqrt{A_i/s_1} \right)$ where π_i is the probability of selecting PSU i and A_i is the area of PSU i. Formula (4) was used since it is much easier to compute than the correct formula. We did, however, compare the two formulae using data from Florida, Georgia, and North Carolina. The results indicated that formula (4) causes a slight overestimate of single-county cost savings and will not alter the conclusions. For derivation of the formula see "Report on the Feasibility of an On-going Program of Interviewer Variance Estimation in the CPS" by Paul Biemer, John Bushery, Ellen Katzoff, Donna Kostanich and Fay Nash. Bureau of the Census 1981.

3 For a more precise definition of principal components see T.W. Anderson (1958), An Introduction to Statistical Analysis, New York, John Wiley and Sons Inc.

DESCRIPTION OF ITEMS USED FOR AHS AND HIS

AHS EVALUATION VARIABLES

OWNER OCC HUs WITH INCOME < $7,000
VACANT YEAR ROUND HUs
OCC HUs WITH < 3 ROOMS
FAMILIES WITH A FEMALE HEAD
OCC HUs WITH 1.01+ PERSONS PER ROOM
OWNER OCC HUs WITH A VALUE < $50,000
RENTER OCC HUs WITH CONTRACT RENT < $200
RENTER OCC HUs THAT PAY 25+% OF INCOME IN GROSS RENT
OWNER OCC HUs BUILT BEFORE 1939
RURAL YEAR ROUND HUs

AHS STRATIFICATION VARIABLES

VACANT HUs FOR RENT
OWNER OCC HUs
OCC MOBILE HOMES OR TRAILERS
OCC HUs LACKING SOME OR ALL PLUMBING
OCC HUs WITH NO COMPLETE KITCHEN FACILITIES
OCC HUs WITH A BLACK HEAD
OCC HUs WITH A HEAD OF SPANISH ORIGIN
URBAN YEAR ROUND HUs
CHANGE IN POPULATION FROM 1970 - 1980
OWNER OCC HUs WITH A VALUE < $25,000
HUs BUILT FROM 1970 TO 1980
HEATING DEGREE DAYS
COOLING DEGREE DAYS

HIS STRATIFICATION VARIABLES

TOTAL UNEMPLOYED
TOTAL SPANISH POPULATION
PERSONS IN URBAN AREAS
PERSONS EMPLOYED IN MANUFACTURING
PERSONS BELOW THE POVERTY LEVEL
OWNER OCC HUs WITH INCOME < $15,000

HIS EVALUATION VARIABLES

TOTAL BLACK POPULATION
PERSONS 60 YEARS OR OLDER
RURAL YEAR ROUND HUs

OCC - OCCUPIED
HU - HOUSING UNIT